A Taxonomy of AI Experiments

Aleksandr Alekseev* Christina Strobel†

University of Regensburg Hamburg University of Technology

June 24, 2025

Abstract

We introduce a taxonomy of artificial intelligence (AI) experiments. Our taxonomy produces four types of AI experiments: conceptual AI experiments, stylized AI experiments, quasi-natural AI experiments, and natural AI experiments. At the core of our taxonomy is the sophistication of AI used, which we evaluate using a simple and robust proxy test of whether AI is developed exclusively for a research study. We discuss the advantages, disadvantages, and best use cases for each type and illustrate the use of each type in various examples. We provide a guide on how to choose the type of AI experiment that best fits a given research question.

Keywords: economic experiments, experimental design, automation, artificial intelligence

JEL Codes: C9, D9, O33

Statements and Declarations: We declare that we have no interests, financial or non-financial, that relate, directly or indirectly, to the research described in this paper.

*Department of Economics, University of Regensburg, Universitätsstr. 31, 93040 Regensburg, Germany, e-mail: aleksandr.alekseev@ur.de

†Institute for Digital Economics, Hamburg University of Technology, Am Schwarzenberg-Campus 4, 21073 Hamburg, Germany, e-mail: christina.strobel@tuhh.de

1 Introduction

The number of experimental papers studying iterations between humans and artificial intelligence (AI) has been steadily rising in recent years (see Figure 1). In fact, almost half of all experimental AI papers, as reviewed in Chugunova and Sele (2022) and March (2021), have been published just since 2013.[1] A number of special issues dedicated to AI research have appeared in various journals such as *Science*, *Management Science*, *Decision Sciences*, *Research Policy*, and *The Journal of Behavioral and Experimental Economics*. With the increasing supply of AI experiments comes a greater demand to organize the literature and make sense of the glut of findings. Recent review articles by Bao et al. (2022), Chugunova and Sele (2022), Langer and Landers (2021), March (2021), Jussupow, Benbasat, and Heinzl (2020), and Burton, Stein, and Jensen (2020), among others, are valuable contributions aimed at meeting this demand. However, recent literature has also identified an unsatisfied need for greater methodological discipline in AI experiments (March, 2021; Langer, König, and Hemsing, 2020). Our paper makes a first step in this direction by introducing a taxonomy of AI experiments and a guide on the best practices for using AI as a tool in experimental economists' methodological arsenal.[2]

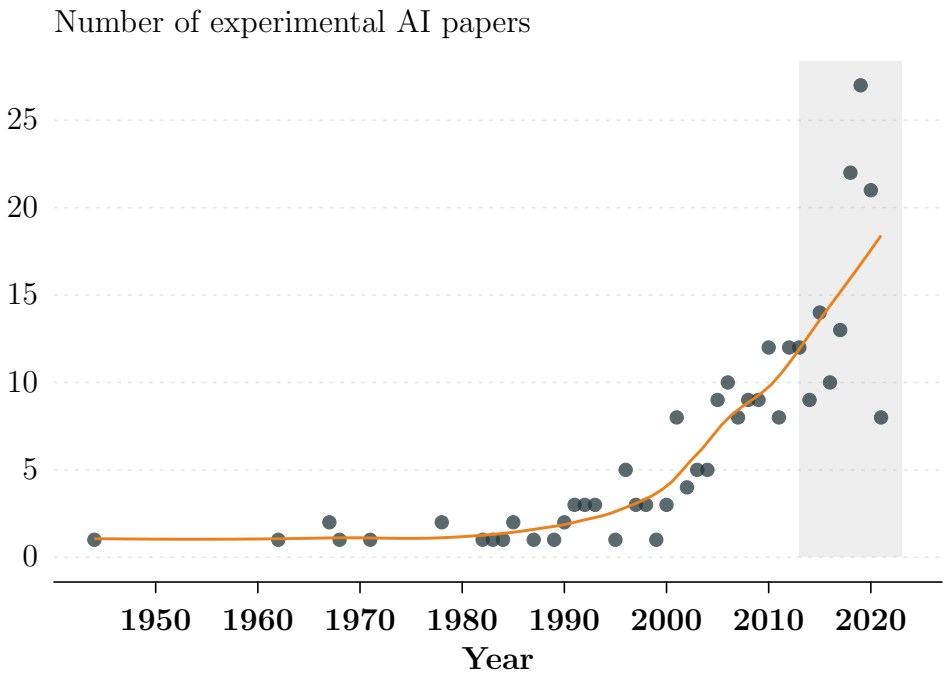
With the growing academic interest in AI experiments, it seems natural to try to define just what *is* an AI experiment. In search of this definition, however, we do not converge on a single ideal type that might be called *the* AI experiment. Rather, we advocate for an inclusive approach and suggest that AI experiments exist on a spectrum that spans quite diverse designs. What defines an AI experiment, in our view, is not so much the design or implementation of AI but rather an underlying research agenda. Therefore, in this paper, we adopt the following broad definition of AI experiments. These are experiments that feature interactions between human subjects and computers, algorithms, artificial agents, machines, robots, automated agents, and artificial intelligence agents with the goal of better understanding how the (actual or hypothetical) interaction with, or the presence of, these agents affects human behavior and outcomes in organizations and markets.[3]

[1] Despite the recent surge in popularity, some of the ideas in current experimental AI research can be traced back as far as the 1950s (Meehl, 1954).

[2] Although the focus of the present paper is on economics experiments, the majority of our discussion applies to other disciplines that conduct AI experiments, including psychology, management, and computer science.

[3] Since the focus of our paper is social sciences, our definition excludes experiments that study exclusively the interactions between AIs. Including them would have strayed us too far into the realm of computer science and AI research. We include robots in this definition and show examples of their use in experiments because robots can be viewed as a physical manifestation of AI (Russell and Norvig, 2021), with the caveat that not all robots are powered by AI.

Figure 1: The growth of experimental AI literature



Note: The figure shows the total number of experimental AI papers published in a given year. The line shows a loess fit. The shaded region shows the time period during which approximately half of all experimental papers were published. We compiled our list based on the papers reviewed in Chugunova and Sele (2022) and March (2021).

Having adopted this broad definition of AI experiments that allows for a variety of designs, we attempt to put some structure on this variety. We identify and label four types of AI experiments: conceptual AI experiments, stylized AI experiments, quasi-natural AI experiments, and natural AI experiments. We discuss the advantages, disadvantages, and best use cases for each type and illustrate their usage in various examples. Our proposed taxonomy can help organize the literature and identify outstanding research questions. For example, a certain phenomenon might have been established only for one type of AI experiments while its robustness in other types of AI experiments is unknown. Our taxonomy can also benefit researchers who are interested in studying AI using experimental methods by helping them settle on designs that best fit their research questions.

At the core of our taxonomy is the sophistication of AI used. The sophistication of AI varies from real-world, commercial-grade systems, such as ChatGPT, to AI-as-label used in vignette studies, and everything in between. Our taxonomy, however, does not rely

on the sophistication of AI directly, which can be subjective and involve dealing with the details of technical implementation of AI that can be impractical to use for classification. Instead, we propose a simple and robust proxy test: whether AI is developed exclusively for the purpose of conducting a research study. The logic behind this proxy test is that real-world, commercial-grade AI systems are unlikely to be developed for the purpose of simply conducting a study. On the other hand, AI developed specifically for an experiment is unlikely to match the sophistication of real-world systems.[4]

We gauge the sophistication of AI in an iterative manner. At a first step, we ask whether AI is just a label, which isolates conceptual experiments. If the AI is not simply a label, we then ask whether it is implemented exclusively for a study, which isolates stylized AI experiments. If AI is not implemented exclusively for a study, we ask if the experiment is conducted in a controlled environment, which isolates quasi-natural AI experiments. Finally, if AI is not implemented exclusively for a study and is used in its natural setting, then we identify it as a natural AI experiment.

Each study should ultimately settle for a type (types) of AI experiment that best fits (fit) the research question at hand.[5] It is possible that no single type best fits a given research question, in which case we recommend a complementary approach. Just like many experimental studies combine different samples (e.g., lab and online) (Hergueux and Jacquemet, 2015; Palan and Schitter, 2018) or types of experiments (e.g., lab and field) (Harrison and List, 2004) that complement each other, AI studies can exploit complementarities offered by different types of AI experiments. For example, a study can first establish a result in a specific setting using a quasi-natural AI experiment that offers greater external validity at the cost of a narrower scope. Then it can explore the mechanisms behind this result in a stylized AI experiment that offers a wider scope at the cost of lower external validity. In any case, we argue that researchers should justify their design choices and provide guidance for such justification.

Our research builds upon and contributes to the recent reviews and methodological discussions of experimental AI papers. The paper that is closest in spirit to the present one is March (2021) that focuses on the use of computer players in experimental games spanning

[4]It is possible that our proxy test can misclassify some cases. For example, a research team conducting an experiment might develop an algorithm that later becomes a commercial-grade system in practice. We believe, however, that the ease of use and robustness of our proxy test outweigh the dangers of potential misclassification that could be caught with a more elaborate test.

[5]It should be noted, however, that while our focus is on experimental studies, some research questions are naturally more amenable to non-experimental methods. We also acknowledge the value of survey-based approaches for studying the use of AI, such as Carvajal, Franco, and Isaksson (2024) and Chugunova et al. (2025).

a range of topics including auctions, bargaining, and social dilemmas. It not only reviews the results from this vast literature but also provides a useful classification of the reasons for using a computer player in a game (e.g., reducing decision-making noise or inducing certain behavioral types) and the types of algorithms used in those games (e.g., equilibrium or adaptive algorithms). Our paper echoes the implicit message in March (2021) that the type of AI used should be based on the research question and addresses the concern about the lack of methodological standardization and guidance in AI experiments.

Bao et al. (2022) offer a more targeted discussion of AI in strategic interactions that occur in experimental financial markets. Their review provides researchers with a detailed classification of the algorithms used in finance experiments and how those algorithms affect participants' behavior and market outcomes. Consistent with our argument that the type of AI experiments should be based on a research question, Bao et al. (2022) highlight the importance of both finance experiments in which the algorithm is actually implemented (our taxonomy would classify the majority of them as stylized AI experiments) and experiments in which the presence of an algorithm is merely announced (these would fall into our conceptual AI experiments category).

Likewise, a review by Chugunova and Sele (2022) embraces methodological diversity. The study stands out due to its breadth and organizes findings from experiments where human subjects interact with automated agents across a wide range of disciplines spanning economics, psychology, sociology, marketing, medicine, and others. While the authors do not make methodological points explicitly, they do note that in many instances the details of AI implementation matter less for subjects' behavior than the mere notion of interacting with AI instead of humans.

Burton, Stein, and Jensen (2020) and Jussupow, Benbasat, and Heinzl (2020) summarize the findings of the literature on algorithm aversion. Although both papers offer illuminating insights into the reasons behind algorithm aversion and potential ways to overcome it, they remain largely silent about the methodological issues in the reviewed studies. Jussupow, Benbasat, and Heinzl (2020) does comment, however, on the predominance of vignette studies, which fall under conceptual AI experiments according to our taxonomy. Perhaps this predominance is what makes the authors refer to AI in a study like Yeomans et al. (2019) as a "real working algorithm," even though our classification puts it into a stylized, rather than a natural or quasi-natural, category.

The methodological concern about the predominance of vignette studies in AI experiments finds a stronger voice in Langer and Landers (2021). This review is unique in its focus on the

so-called second, i.e., people who are affected by AI but do not interact with it, and third-party, i.e., outside observers, behavioral reactions to AI rather than on the more common first-party, i.e., people who interact with AI, behavior. Langer and Landers (2021) lament the over-reliance on vignette studies, which have the downside of lower external validity. Our paper acknowledges this important shortcoming of conceptual AI experiments, which include vignettes, however, we also highlight their benefits and discuss those shortcomings and benefits in relation to other types of AI experiments. Importantly, we do not take a stance on what type of AI experiments is "best" but instead argue that the type of AI experiment should fit the research question.

Our study is also related to the literature that explores the broader methodological role of AI in scientific research and experimentation. Charness, Jabarian, and List (2025) offers a related yet distinct perspective by focusing on the application of generative AI, in particular large language models (LLMs), as a research tool. The authors argue that LLMs can enhance experimental research by improving comprehension, immersion, data collection, and analysis. The review also addresses broader risks and benefits, providing guidance on how generative AI might support open science and enable scalable experimentation in policy and business contexts. In a similar vein, Korinek (2023) investigates the potential of generative AI, such as ChatGPT, to support economists. The study identifies six key areas where generative AI can be beneficial: ideation, writing, background research, data analysis, coding, and mathematical derivations. While these reviews offer practical guidance and examples showcasing the potential of LLMs throughout the research process, our study focuses on how AI is implemented within experiments and explores the diverse approaches for designing AI experiments.

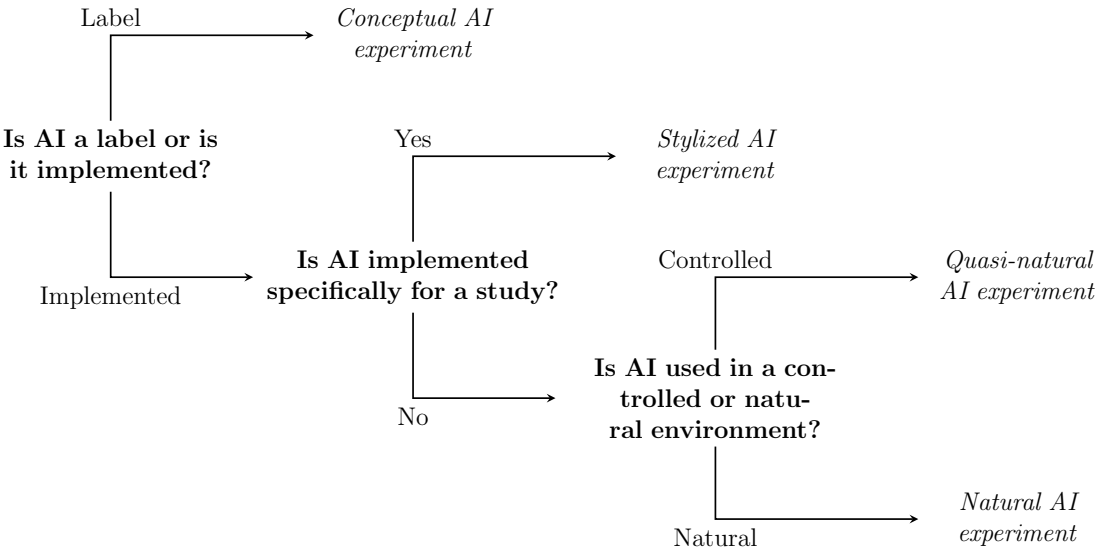
Our paper offers three main contributions to the literature. First, we propose a taxonomy of AI experiments and a simple and robust test to classify studies according to this taxonomy. Second, we discuss the advantages and disadvantages of each type of AI experiments that we identify and compare the types based on these features. Third, we discuss the best use cases for each type along with relevant examples from the literature. Our paper, however, is not a substitute for existing taxonomies, e.g., March (2021) or Bao et al. (2022). Rather, we complement these by providing researchers with a high-level classification of AI experiments and guidance on how to choose the right type to fit their research questions, after which researchers should tailor their designs using narrower taxonomies developed for their topics. We also note that our paper is not a literature review. Our examples of studies using AI experiments are not meant to be exhaustive. Instead, they are meant to illustrate why choosing a particular type of AI experiment makes sense in the context of these studies.

2 A Taxonomy of AI Experiments

2.1 Classification Procedure

We identify and label four types of AI experiments: *conceptual AI experiments*, *stylized AI experiments*, *quasi-natural AI experiments*, and *natural AI experiments*. Figure 2 shows the decision tree that we use for classification. Conceptual AI experiments, unlike the three other types, are characterized by AI that exists merely as a label or a framing device. In stylized AI experiments, the AI is implemented but is designed specifically for a study. In both quasi-natural and natural AI experiments, the AI used in a study is designed for purposes other than conducting a study. What distinguishes these two types is the environment in which the use of AI occurs. In quasi-natural AI experiments, the use of AI occurs in a controlled environment. Natural AI experiments, on the other hand, take place in an environment in which the use of that AI naturally occurs.[6]

Figure 2: Decision tree for classification



[6]We use the following rationale for our terminology. Natural AI experiments, by analogy with the natural field experiments, occur in an environment that is "natural" for the AI used in them, hence the term. Quasi-natural AI experiments are "almost like" the natural AI experiments since both use the same type of AI, hence the term "quasi." The only difference between the two is the environment in which experiments occur, which we view as minor albeit important. We use the term "stylized" in its dictionary definition sense ("depicted or treated in a mannered and nonrealistic style"), and similarly to "stylized facts." "Stylized" here refers to AI that is deliberately nonrealistic and removed from many real-world implementation details in favor of focusing only on the essential elements. Finally, "conceptual" refers to AI as a concept or an idea rather than an actually implemented device. We do not call conceptual AI experiments "quasi-stylized" because in our view the difference between the conceptual and stylized AI experiments is more substantial than the difference between natural and quasi-natural AI experiments. Conceptual AI experiments have no implementation of AI whatsoever, while stylized AI experiments do.

Although we present these types in a certain order, we stress that the order does not reflect their quality. While it might be tempting to claim that one type of AI experiments is superior to another, we argue that this is not the case. Each type of AI experiments has its advantages and disadvantages, and there is no single type that is suitable for all studies. We now proceed to define each type, discussing its advantages, disadvantages, and best use cases. Table 1 offers an overview of each type.

Table 1: Overview of advantages and disadvantages of each type

Type	**External Validity**	**Ease of Implementation**	**Replicability**	**Control over Algorithm**	**Scope of Questions**	**Best use cases**
Conceptual	*	****	****	****	****	Discovery-oriented
Stylized	**	***	***	***	***	Complex, dynamic interactions
Quasi-natural	***	**	**	**	**	Use-inspired basic research
Natural	****	*	*	*	*	Practice-focused

2.2 Conceptual AI Experiments

In conceptual AI experiments, AI exists merely as a label or framing device that models the operational principles or consequences of AI, but no AI is actually implemented. These experiments typically take place in a controlled, rather than natural, environment. Not being constrained by the implementation of AI allows researchers to ask questions and present scenarios, such as ethical dilemmas or changes in a labor market, that would be impractical or even impossible to test with actual AI. A primary advantage of conceptual AI experiments is that they are relatively inexpensive. Researchers can conduct them at scale and replicate them easily. The low cost of these experiments allows researchers to quickly explore mechanisms underlying subjects' behavior or test the generalizability of their findings across different subject pools. The downside of conceptual AI experiments is that subjects' choices are typically inconsequential, which may affect the external validity of results in environments with real stakes. The best use case for conceptual AI experiments is when researchers wish to make a first attempt at answering a new research question or to complement other AI designs.

2.3 Stylized AI Experiments

In stylized AI experiments, AI is actually implemented, unlike in conceptual AI experiments, however, it is designed specifically for a research study. Stylized AI experiments are conducted in a controlled, rather than natural, environment because the AI used in them typically does not occur outside the study. AI used in stylized AI experiments is based on a

research question and thus can take a variety of forms, ranging from rule-based algorithms to algorithms that replicate decisions based on historical data and artificial agents trained using reinforcement learning.

The major strength of stylized AI experiments is that they can help answer a broad range of questions about human-AI interaction in a setting with real stakes. These experiments are not constrained by AI systems that already exist, i.e., a research question comes first and AI comes second. Another benefit of these experiments is that they are feasible (i.e., they can be conducted with standard lab or online samples) and replicable. The downside of stylized AI experiments is lower external validity relative to natural and quasi-natural AI experiments due to their reliance on AI that is unlikely to be used outside a study. The best use case for stylized AI experiments is when a research question cannot be answered using an existing AI system, yet a study requires some implementation of AI. For example, an experiment might feature complex and dynamic interactions between players that conceptual AI experiments cannot capture.

2.4 Quasi-Natural AI Experiments

Quasi-natural AI experiments feature realistic AI that is designed for purposes other than a research study, unlike in stylized AI experiments. Quasi-natural AI experiments occur in a controlled, rather than in a naturally occurring, environment—a feature that they share with conceptual and stylized AI experiments. An example of a quasi-natural AI experiment is studying how human subjects interact with a state-of-the-art chatbot, such as ChatGPT, or commercial-grade robots in a controlled, albeit artificial, environment of an experimental lab.[7] Quasi-natural AI experiments can also occur when an organization runs pilot experiments, e.g., to test which features of AI lead to better customer satisfaction, before launching a product.

The main strengths of quasi-natural AI experiments are the realism and external validity that they inherit from natural AI experiments while being much more feasible, affordable, and replicable. Quasi-natural AI experiments can be used to study a broad set of research questions because these experiments are not tied to the natural environments in which the use of AI occurs. Their affordability also makes them a convenient testbed before scaling up a study. The reliance on actual AI, however, can be a limitation. Quasi-natural AI experiments use AI that is not designed specifically for a research study, i.e., AI comes first, research comes second, hence, the set of research questions that such experiments can answer

[7]To provide a counter-example, if the developer of ChatGPT, OpenAI, conducts an experiment on the website that people use to access ChatGPT, that would count as a natural AI experiment.

is constrained by AI itself and is a priori smaller than for conceptual or stylized experiments. The best use case for quasi-natural AI experiments is when a realistic implementation of AI matters, but it is unaffordable, impractical, or unethical to conduct a natural AI experiment.

2.5 Natural AI Experiments

Natural AI experiments use sophisticated AI—such as computer vision, natural-language processing, decision support, recommender systems, and robots—in environments in which the use of that AI naturally occurs. The AI is usually trained on Big Data using methods that transcend simple rule-based approaches and encompass a variety of supervised or unsupervised machine learning techniques. The defining characteristic of natural AI experiments is that the AI involved is developed for purposes beyond the scope of the research study itself. In natural AI experiments, AI comes first, and a research study comes second. Natural AI experiments are also natural field experiments (Harrison and List, 2004).

Many natural AI experiments take the form of the so-called *A/B testing* (Azevedo et al., 2020) that tech companies such as Microsoft, Google, Amazon, Spotify, and Netflix run on their platforms to increase engagement or develop product innovations. In these experiments, AI is not designed for the purpose of conducting a research study but, for example, for the purpose of giving better recommendations to the users of a service or improving the search results on a search engine, which ultimately benefit an organization. These experiments occur on a platform itself during the actual use of the services by its users.

The benefit of natural AI experiments is that they possess the highest possible degree of realism and external validity. The results of these experiments are directly applicable to the research questions asked in a study. The flip side of that realism and the most prominent downside of natural AI experiments is that they are often too narrow and hard to generalize outside the context of an experiment. Tech companies run thousands of A/B testing experiments, often automated, on their platforms, but only a handful of them produce generalizable knowledge about human-AI interactions. Another downside of natural AI experiments is low feasibility and replicability. Natural AI experiments are rarely public, and even the ones that are made public would be difficult to replicate because they would involve hard-to-obtain access to an organization's proprietary platform. The best use case for a natural AI experiment is when its results are both useful to the organization where it is conducted and broad enough to contribute to general knowledge about human-AI interaction.

3 The Taxonomy in Action

In this section, we put our taxonomy into action and illustrate the use cases of each type through various examples. Table 2 provides a summary of the reviewed papers.

Table 2: Summary of reviewed papers

Classification	Experiment	Topic	AI implementation
Conceptual AI Experiments	Awad et al. (2018)	moral beavior	vignettes
	Lee (2018)	acceptance	vignettes
	Castelo, Bos, and Lehmann (2019)	acceptance	vignettes
	Granulo, Fuchs, and Puntoni (2021)	acceptance	vignettes
	Wu (2022)	preferences	vignettes
	Zhang (2022)	preferences	vignettes
	Gallego et al. (2022)	preferences	vignettes
	Jeffrey (2021)	preferences	vignettes
	Farjam and Kirchkamp (2018)	behavior	rule-based
	Angerer, Neugebauer, and Shachat (2023)	behavior	rule-based
	Jacob Leal and Hanaki (2023)	behavior	rule-based
Stylized AI Experiments	Strobel (2025)	performance	timing
	Alekseev (2025)	preferences	rule-based
	Kirchkamp and Strobel (2019)	altruism	historical averages
	Corgnet, Hernán-González, and Mateo (2023)	social pressure	historical averages
	Gogoll and Uhl (2018)	preferences	rule-based
	Dietvorst, Simmons, and Massey (2015)	preferences	regression
	Dargnies, Hakimov, and Kübler (2024)	preferences	regression
	Klockmann, von Schenk, and Villeval (2022)	altruism	supervised learning
	Werner (2021)	cooperation	reinforcement-learning
	Schauer and Schnurr (2023)	cooperation	reinforcement-learning
	Dell'Acqua (2022)	performance	supervised learning
Quasi-Natural AI Experiments	Cominelli et al. (2021)	trust	image recognition
	Gorny, Renner, and Schäfer (2023)	altriusm	robot
	Leib et al. (2023)	moral behavior	GPT-J
	Dell'Acqua et al. (2023)	performance	GPT-4
	Cox et al. (2016)	performance	probit model
Natural AI Experiments	Luo et al. (2019)	performance	chatbot
	Brynjolfsson, Li, and Raymond (2025)	performane	GPT-3
	Paravisini and Schoar (2013)	financial decisions	scoring model
	Bundorf, Polyakova, and Tai-Seale (2019)	acceptance	scoring model
	Bai et al. (2022)	preferences	rule-based

3.1 Conceptual AI Experiments

The most common application for conceptual AI experiments is vignette studies. A typical vignette study presents subjects with a series of hypothetical situations and asks them to state their preferences about what they would do in those situations. The primary advantage of vignette studies is that they can model any situation of interest to researchers without being constrained by the actual implementation of AI.

For example, Awad et al. (2018) conduct an online experiment that elicits subjects' preferences for what a self-driving car should do in moral dilemmas (trolley problems). A typical

dilemma involves a hypothetical situation with a malfunctioning self-driving car that can either stay on course and kill pedestrians or swerve and kill the passengers. The dilemmas present subjects with various trade-offs such as between humans vs. animals, more lives vs. fewer lives, and young vs. old, and each subject makes choices in 13 different situations. Using vignettes allows the authors to collect an enormous dataset of choices spanning a wide variety of situations and cultures. Moreover, a conceptual AI experiment is the most appropriate choice for this research question. Running a natural AI experiment would have been infeasible: One could never run an experiment where actual self-driving cars implemented subjects' choices and killed people. One could attempt to make subjects' choices consequential by promising to use these choices to inform the programming of actual self-driving cars in the future. The legal or ethical status of such a promise, however, would be unclear. A more feasible alternative would be a stylized AI experiment in the spirit of the "mouse-model" design of Falk, Neuber, and Szech (2020). Researchers could program simple devices that would implement subjects' choices that would be consequential for mice. Aside from the ethical issues it might raise, the downside of this implementation is that mice do not possess many of the characteristics of interest of potential victims in Awad et al. (2018).

AI vignettes often present subjects with the option to choose between an AI and a human. For example, Lee (2018) presents subjects with different managerial decisions and asks them about their perceptions of those decisions when implemented by either AI or a human. Castelo, Bos, and Lehmann (2019) offer subjects a variety of tasks and ask them whether they prefer an AI or a human to complete those tasks. Similarly, Granulo, Fuchs, and Puntoni (2021) ask subjects whether they prefer an AI or another human to replace a human worker. In addition to exploring a variety of choice situations, these studies highlight two other key advantages of conceptual AI experiments. First, these experiments allow researchers to quickly explore the mechanisms underlying subjects' behavior, for example, whether a decision requires mechanical or human skills (Lee, 2018), whether a task is objective or subjective (Castelo, Bos, and Lehmann, 2019), and whether the replaced worker is the subject herself or a third party (Granulo, Fuchs, and Puntoni, 2021). Second, researchers can quickly test the replicability and generalizability of their results across different subject pools. Unlike the Awad et al. (2018) study, the studies cited above could be feasibly run as natural, quasi-natural, or stylized AI experiments. Some scenarios, such as the work assignment scenario in Lee (2018), have even been implemented as natural AI experiments (Bai et al., 2022). However, even the studies that use natural AI implementations, such as that by Bai et al. (2022), acknowledge the complementary value of conceptual AI experiments.

Another useful application for conceptual AI experiments is the manipulation of subjects'

beliefs through priming. For example, the political-science experiments by Wu (2022), Zhang (2022), Gallego et al. (2022), and Jeffrey (2021) use priming to evaluate the effects of subjects' beliefs about their automation exposure on their tendency to support various public policies. In a typical priming experiment, subjects first read a passage containing some information about new technologies, which is supposed to affect their beliefs, or some neutral information, which serves as a control condition, and then have to state their support for given policies. The conceptual AI experiments are ideal for these studies because researchers are typically interested in the general effects of automation. Running a natural AI experiment in this context would be challenging because it is difficult to experimentally manipulate the automation exposure for actual workers.[8] Running a stylized or quasi-natural AI experiment would be an option, and in fact, experiments that elicit distributional preferences in controlled environments are not uncommon in economics (Cappelen, Cappelen, and Tungodden, 2023).

Conceptual AI experiments are not limited to vignette studies. A series of finance experiments by Farjam and Kirchkamp (2018), Angerer, Neugebauer, and Shachat (2023), and Jacob Leal and Hanaki (2023) leverage the possibility of interacting with AI to study actual changes in subjects' behavior. In these experiments, subjects participate in stylized financial markets in which they can buy and sell assets for several rounds. The treatment condition informs subjects that the markets in which they participate may include algorithmic traders. The control condition does not give subjects this information. Importantly, no algorithmic traders exist in either condition.[9] These studies employ conceptual AI experiments mainly to disentangle the mechanisms underlying subjects' behavior. The presence of algorithmic traders in a market can affect the behavior of human traders and the resulting market outcomes through two distinct channels. The first channel is mechanical: The algorithms implement certain strategies and trade faster, which directly influences market outcomes. The second channel is behavioral: The beliefs about the presence of algorithms and what strategies they employ may alter the behavior of human traders independently of what the algorithms actually do. By focusing on the possibility of interacting with algorithmic traders, the studies cited above are able to isolate the behavioral channel. None of the other types of

[8] Although experimental methods allow for a more precise way of manipulating and measuring beliefs, in this case a more feasible alternative to a natural AI experiment could be an observational study that uses an exogenous shock as an instrument for automation exposure (Anelli, Colantone, and Stanig, 2019; Webb, 2019).

[9] Unlike some AI experiments in computer science and psychology, economic and finance experiments do not use deception. Even though the studies cited here do not focus on the treatments that feature interactions with algorithmic traders, there was an actual possibility of interacting with algorithms in each study. Angerer, Neugebauer, and Shachat (2023) is an exception: It does discuss treatments that implement algorithmic traders.

AI experiments would have been suitable for achieving this research goal because the actual implementation of algorithmic traders would have confounded the two channels.

3.2 Stylized AI Experiments

Unlike conceptual AI experiments, stylized AI experiments actually implement AI in some form. The ability to design their own algorithms presents researchers with a variety of implementation options. One of the simplest options is a rule-based algorithm. Such algorithms are easy to explain to subjects and offer a high degree of control over the performance of the algorithm.

For example, Strobel (2025) studies whether automated bonus evaluation affects worker performance using a modified one-shot principal-agent game. In the game, subjects in the role of workers choose their performance levels. Subjects in the role of principals set performance thresholds for assigning bonuses. The performance threshold can either be set before the performance is known, which models the mechanism of an automated bonus evaluation process, or after the performance is known, which models the mechanism of a non-automated process. The automated process is implemented by simply comparing the predetermined performance threshold with the actual performance and assigning a bonus if the threshold is exceeded, and subjects are aware of each process. The experiment varies whether the bonus evaluation process is determined by a principal or randomly. The results show that performance is significantly lower under the automated process. However, whether automation is determined by a principal or randomly has no significant effect on performance. The study argues that lower performance under the automated process is not driven by fairness or trust concerns, but rather by misaligned expectations about how generous the threshold should be.

Alekseev (2025) is another example of a rule-based algorithm in a labor setting. It studies preferences for working with an algorithm that induces a task-switching environment. In the experiment, subjects perform real-effort tasks and make a choice of whether to work manually and complete the tasks themselves or work with an algorithm and delegate some of the tasks to it. Delegating to the algorithm enables subjects to work on new tasks, which is always better than completing all the tasks themselves in terms of monetary payoffs. The algorithm, however, is programmed to periodically interrupt subjects' work and ask for help. Subjects are aware of the interruption rule—the algorithm interrupts only when it encounters certain tasks, the frequency of which is identical whether subjects work by themselves or with an algorithm—but not of the actual interruption frequency, which varies between subjects. The study finds that as the frequency of interruptions increases, subjects are less likely to

delegate to the algorithm, which suggests that task switching has tangible utility costs to the subjects.

An alternative to rule-based algorithms are algorithms that are trained on past data. A common way to train such algorithms is to simply replicate the distribution of decisions from past human-only sessions. Automating the decisions of other players with such AI allows researchers to isolate the behavioral effects of mechanisms such as social preferences, intentionality, or peer pressure (March, 2021). By replacing a human player with an AI that plays like an average human while not possessing the relevant characteristics of a human, researchers can shut down the mechanisms of interest and see how it affects the behavior of human subjects.

An example of a stylized AI experiment that uses historical averages to automate decisions is Kirchkamp and Strobel (2019). The study investigates the role of the perceived responsibility and guilt of others on one's own responsibility, guilt, and selfish choices. In the experiment, human subjects play a one-shot dictator game in which the dictator's decision to split money equally or unequally is implemented by a pair of players. The pairs of players consist of either two human players or a human player and a passive human player whose decisions are automated by an algorithm. Human subjects in the latter case know that a computer automates the decisions of a passive player by replicating past choice frequencies but not the actual frequencies. The researchers expect human dictators to feel more responsible for an outcome, feel more guilt for an unequal split, and make fewer selfish choices when they are paired with an algorithm than with another human because an algorithm cannot be responsible and feel guilt in the same way a human does. The researchers, however, find no such effects.

Corgnet, Hernán-González, and Mateo (2023) is another example of using historical averages, this time in a labor setting, that examines the effects of social pressure on workers' performance. In the experiment, a team of three workers repeatedly performs, over five rounds, a sequential task mimicking an assembly line. The team consists of either only human workers or two human workers and one algorithm. The researchers calibrate the productivity of the algorithm to be the same as that of an average human worker. Subjects working with an algorithm are aware neither of the actual productivity of an algorithm nor of the calibration rule, only of the existence of an algorithm, however, they can observe the algorithm's performance. The researchers isolate the social-pressure effect because the algorithm cannot impose social pressure on other workers in the same way a human can. The study finds that subjects who work with the algorithm underperform relative to those who work in human-only teams, which highlights the importance of social pressure for team

performance.

Gogoll and Uhl (2018) is an example of using historical averages to train AI in an experiment at the intersection of labor and moral domains. It studies preferences for delegating to an algorithm tasks that affect third parties. In the experiment, subjects make a choice of whether to delegate a numerical task to another human subject or to an algorithm. The performance on the task, however, affects not the payoff of the subject who solves the task but that of another subject, adding a moral component to the delegation choice. The algorithm is programmed to reproduce the performance distribution of human subjects in previous sessions. As in Corgnet, Hernán-González, and Mateo (2023), subjects are aware neither of the actual performance of the algorithm nor of the calibration rule, however, they observe a snapshot of the algorithm's performance (along with that of human subjects) before they make their delegation decisions. The study finds that subjects are three times more likely to delegate the task to another human than to an algorithm, despite identical ex-ante performance. The researchers find that neither the perceived differences in performance between an algorithm and humans nor trust in the algorithm can explain the reluctance to delegate to an algorithm.

A more sophisticated approach for training AI on past data is to estimate a prediction model, such as an ordinary least squares (OLS) regression. This approach provides less experimental control over the performance of the algorithm, which is typically better than that of human subjects, and is more difficult to explain to subjects. However, the superior performance of such algorithms is often a desirable feature that they share with commercial-grade AI tools, which ultimately increases the external validity of a study.

For example, Dietvorst, Simmons, and Massey (2015) study subjects' preferences for performing a forecasting task themselves or delegating the task to an algorithm. In the experiment, subjects have to predict over ten rounds how successful an MBA student would be using such variables as a student's undergraduate degree, GMAT scores, years of work experience, and education. The subjects can either make all predictions themselves or delegate all predictions to an algorithm. To build an algorithm, the researchers estimate an OLS regression on the data from 115 students using the same explanatory variables as the subjects themselves can use. The instructions give subjects general information about the algorithm ("The model is based on hundreds of past students, using the same categories of demographic data you are receiving") but not the actual implementation. Even though the algorithm is better than humans at prediction, a significant fraction of subjects does not choose the algorithm. The study additionally proposes and explores a potential mechanism behind such algorithm aversion: Observing an algorithm perform may increase aversion to

it. Consistent with this hypothesis, the fraction of subjects who choose an algorithm drops if the subjects observe the algorithm's performance.

Dargnies, Hakimov, and Kübler (2024) use a similar approach to study the preferences of workers and managers for using an algorithm to evaluate workers and make hiring decisions. In the experiment, subjects in the role of workers perform real-effort tasks. They then make a choice of whether they prefer the hiring decision between themselves and another worker be made by another human subject in the role of a manager or by an algorithm. Subjects in the role of managers first make 20 hiring decisions from among pairs of workers and then make a choice for whether they want to delegate their hiring decisions to the algorithm. The managers can use workers' task performance and gender to make their hiring decisions. To build an algorithm, the researchers estimate an OLS regression on the data from 200 workers using the same explanatory variables that are available to managers (i.e., task performance and gender). The subjects know that the algorithm is designed to predict performance based on the data from previous workers, and that it hires the worker with the highest predicted performance, while the information about implementation details varies by treatment. The study finds that both the workers and the managers prefer human evaluation over algorithmic evaluation, despite the algorithm being better at picking the better-performing worker. However, when the algorithm does not use a worker's gender for prediction and workers know this, they choose the algorithm more often. The study additionally finds that explaining how the algorithm works does not increase either workers' or managers' preference for it.

Researchers are not limited to using simple prediction models, such as OLS. However, more sophisticated models, while offering better predictions, also require better explanations. For example, Klockmann, von Schenk, and Villeval (2022) study how subjects' behavior is affected if they know that their choices train an algorithm that later makes a decision that has consequences either for them or for other subjects. In the experiment, subjects in the role of dictators first make repeated choices in a dictator game for 30 periods. The researchers then create an algorithm for each dictator to predict and make the choice in the final 31st period. To create the algorithm, the researchers train a random forest model using such features as the payoffs and the sum and difference of points allocated to a receiver in the dictator game. The instructions inform subjects that the dictators' choices are used to "train an artificially intelligent Random Forest algorithm," and give a summary of how the algorithm works. The prediction of the algorithm is implemented for either a receiver with whom a dictator was paired, a receiver in a different pair, or, with some probability, the dictators themselves. The study finds that the behavior of dictators does not differ between the cases when the

algorithm makes a decision for a receiver in the dictator's own pair or for a receiver in a different pair. However, if there is a chance that the algorithm determines the payoff of the dictators, the dictators behave more prosocially: The share of egalitarian decisions increases.

Yet another step in the complexity of algorithms used in stylized AI experiments is artificial agents trained using reinforcement-learning. Such agents can be used to study complex decisions in dynamic strategic environments. A popular reinforcement-learning algorithm in experimental research is Q-learning. For example, Werner (2021) and Schauer and Schnurr (2023) study market outcomes and strategies in experimental oligopoly markets populated by either human participants, Q-learning agents, or both. In both studies, subjects are aware of whether they are playing against other humans or algorithms while the strategy of algorithms is not disclosed.[10] Werner (2021) additionally informs subjects that an algorithm "acts in the interest of another participant" who does not make any decisions but receives profits earned by the algorithm. The results show that both humans and algorithms learn to collude with each other, even without communication. However, collusion is highest when only human or only Q-learning agents are present in a market, while in hybrid markets humans and algorithms fail to coordinate.

In all the above examples, stylized AI experiments are the most fitting design choice for answering the posed research questions. This design allows researchers to create custom AIs that are tailored to their research questions and to easily evaluate different potential mechanisms behind subjects' choices. Conceptual AI experiments cannot capture the consequential nature of decisions often desired, e.g., for moral choices, or the dynamic choice environment of market experiments. Quasi-natural AI experiments would have been too restrictive in the types of algorithms researchers can use because commercial-grade AI systems often do not exist for the types of questions explored in these studies. Natural AI experiments, on the other hand, would be challenging to implement because these studies are typically interested in general patterns of human behavior without immediate benefits to an organization. For some research questions, e.g., the ones that are motivated by labor settings that are of interest to organizations, stylized AI experiments provide a convenient first step for testing ideas before scaling them up to a level of natural AI experiments.

3.3 Quasi-Natural AI Experiments

Quasi-natural AI experiments, similar to stylized ones, actually implement AI in a controlled environment. The AI in them, however, is designed for purposes other than a research study.

[10] Schauer and Schnurr (2023) study has one treatment where subjects do not know the identity of their AI competitor, while the other four treatments reveal this information.

An example of a quasi-natural AI experiment is Cominelli et al. (2021) who employ a robot to study the effects of promises made by a robot or a human on trust in human counterparts in one-shot games. The researchers use the Facial Automaton for Conveying Emotions (FACE) robot designed for social robotics and, in particular, therapy for autism (Pioggia et al., 2004). The FACE robot has a human-like appearance and is capable of showing emotional states, empathy, and nonverbal communication. In the experiment, a human subject enters a room containing either a robot, a professional actress, or a computer, each of which makes a verbal promise to take a cooperative action. After that, the subject decides whether to trust them. The study finds that receiving a promise from the robot increases trust in participants who perceive the robot as human-like, but not in those who do not perceive the robot as such. A similar pattern occurs when either an actress or a non-human-like computer makes a promise. Using a quasi-natural AI experiment in this study strikes a perfect balance between producing generalizable knowledge and realism. On the one hand, a natural AI experiment would not have been practical because the researchers are interested in a general pattern of human behavior, not in something that could directly benefit an organization. On the other hand, researchers are interested in subjects' responses to a human-like robot, so a stylized or a conceptual AI experiment would have been too abstract.

Gorny, Renner, and Schäfer (2023) is another example of a quasi-natural AI experiment that uses state-of-the-art robotics. The researchers conduct a lab experiment at a learning factory to study how the presence of robots in a team of human workers affects their prosociality towards each other and the valuation of their products. A learning factory is a real, albeit simplified, production system featuring state-of-the-art robotics where university employees and students receive hands-on training on production technologies. In the experiment, two human workers produce electronic motor components by operating two different production stations at the beginning and end of a three-station production line. The middle station is operated by either two robots or a station that performs the same steps but with the robots switched off and hidden. The study finds that the presence of robots increases sharing behavior among human workers, however, it does not change workers' valuation of the rewards they earn from production. A quasi-natural AI design allows the researchers to achieve the right balance between realism, on the one hand, and generalizability and affordability, on the other. The researchers are interested in subjects' behavior in a realistic production environment where subjects have no ambiguity about what a robot is and are engaged in a meaningful production task. A conceptual or a stylized design would have been too simplistic for that purpose. A quasi-natural design, however, is an affordable testbed before scaling the study up to the level of a natural AI experiment.

Leib et al. (2023) is another example of a quasi-natural AI experiment, but it uses a large language model to study how AI- versus human-generated advice affects dishonest behavior among human participants in a one-shot die-rolling game. The experiment employs the GPT-J model developed by EleutherAI in 2021 (Wang and Komatsuzaki, 2021). The researchers fine-tune the model using text advice generated by human participants. In the experiment, subjects observe either AI-generated advice, human-generated advice, or no advice and then decide whether to take a dishonest but self-serving action. The instructions explained to subjects the basics of what the language model is and how it was trained. The results show that advice that promotes dishonesty increases dishonest behavior, while advice that promotes honesty does not increase honest behavior. This pattern holds regardless of whether an AI or a human gave that advice and regardless of whether a subject knows the exact source of the advice. The natural AI design would not have fit the purpose of this study because the researchers are interested in a general pattern of human behavior that has no direct relevance for an organization's profits. A conceptual AI design also would not have been the right fit because the researchers are interested in subjects' actual behavior in response to actual advice. Hence, either a stylized or quasi-natural AI design would have been a good fit. The authors ultimately opt for the quasi-natural design to increase the external validity of their findings.

Dell'Acqua et al. (2023) is another example of a quasi-natural AI experiment, this time within a consulting sector, that studies the effect of access to a generative AI tool among highly skilled knowledge workers on their productivity and quality of work. The researchers conduct an experiment with the consultants at Boston Consulting Group, a leading consulting firm, who are randomly assigned to either use the AI tool with no guidance, the AI tool with prompt-engineering training, or no AI at all. The AI tool in the experiment is the GPT-4 model by OpenAI. In the experiment, subjects complete a series of stylized tasks, representative of typical consulting activities at the company, and are then scored on the quality of their responses. The tasks differ by whether they are within the "AI capability frontier" (tasks that AI can reliably perform) or not (tasks that AI cannot reliably perform). The study finds that for tasks inside the frontier, access to the AI tool increases the number of completed tasks by 12.2% and quality by over 40%. These effects are most pronounced among lower-performing subjects, suggesting that AI acts as a performance equalizer. On the other hand, for tasks outside the frontier, access to AI reduced accuracy by 19 percentage points. Since the researchers are interested in the effect of a commercial-grade AI tool on the behavior of actual consultants, a conceptual or a stylized design would have been too simplistic for that question. A quasi-natural design then acts as a convenient testbed before scaling the study up to the level of a natural AI experiment.

3.4 Natural AI Experiments

Natural AI experiments are typically conducted in collaboration with an organization that is interested in deploying a new technology on its platform. For example, Luo et al. (2019) conduct a field experiment with a large Chinese internet-based financial-services company. The researchers study the effects of using a chatbot on the outcomes of sales calls to the company's customers who are eligible for loan extensions. The company uses a sophisticated voice AI chatbot trained on the voice data of the best-performing human workers that can conduct natural-sounding conversations indistinguishable from human conversations. The study finds that the chatbot is as effective in making sales as the best human workers and four times more effective than inexperienced workers. However, revealing that the caller is a chatbot reduces sales by about 80%, a drop mostly driven by customers' biases against machines. Using a natural AI experiment is ideal in this scenario. On the one hand, the experiment informs the company about how to increase sales and reduce costs: By revealing the identity of a chatbot after a sale is made. On the other hand, the experiment produces knowledge about human-AI interactions: Humans are reluctant to make a purchase solicited by a chatbot, even if it is barely distinguishable from a human seller conveying the same sales pitch. Although other types of AI experiments could produce similar general insights about human-AI interaction, only a natural AI experiment offers the ultimate test of how a company's actual customers would respond to a chatbot.

Brynjolfsson, Li, and Raymond (2025) presents another example of using AI in a customer-support setting. The authors collaborate with a Fortune 500 company that sells business-process software to study the effects of the deployment of an AI assistant among over 5,000 customer support workers on their productivity and service quality. The AI assistant is based on the GPT-3 model by OpenAI and is designed to monitor conversations in real-time and provide suggested responses and links to technical documentation. The design of the AI assistant allows for worker discretion over whether to follow its recommendations. The study finds that access to the AI assistant leads to 15% more issues resolved per hour, the study's main measure of productivity. The authors also report treatment effect heterogeneity: less experienced and lower-skilled workers benefit the most, while skilled workers see marginal speed gains. The results suggest that generative AI functions by transferring tacit knowledge from high-performing agents to less experienced ones, effectively leveling up the baseline capability of the workforce. A natural AI experiment is the most fitting choice in this scenario. On the one hand, the experiment informs the company about the effects of introducing a new technology. On the other hand, it produces knowledge about human-AI interactions: the AI tools impact workers differently, depending on their skill level. While

other types of AI experiments could have been used to detect similar effects, only a natural AI design informs the company about the actual productivity gains of the new technology.

Paravisini and Schoar (2013) is a further example of a natural AI experiment but in the banking sector. The researchers conduct a field experiment with a for-profit bank in Colombia to examine the impact of a credit-scoring model used to evaluate prospective borrowers on the behavior of loan officers. The study finds that the committees that observe credit scores spend more time per loan application and reach decisions more often than the committees that do not observe the scores. The percentage of non-decisions also drops by more than 40%, with the effect being concentrated in difficult-to-evaluate applications. Interestingly, the committees who observe the scores only after making interim decisions also increase their output, by 75%, even though observing the scores never changes their interim decisions, and the quality of decisions is similar to that of committees who observe the scores before making their decisions, which suggests an incentive effect of scores availability. The bank can benefit from these findings by incorporating their credit-scoring system into the loan-approval process, and a follow-up study can help determine whether the score should be presented before or after a decision is made. On the other hand, these findings show how groups of human workers respond to the availability of an algorithmic score: They work harder and reach decisions more often, which suggests a coordinating role of a score. As before, other types of AI experiments could have revealed the effect of a score's availability on the performance of groups of workers, but only a natural AI experiment achieves that in a context that matters for the company.

Although natural AI experiments often benefit an organization's bottom line, as the previous examples illustrate, sometimes they are used for helping customers or improving the overall efficiency of a system. A case in point is Bundorf, Polyakova, and Tai-Seale (2019) who conduct a field experiment in collaboration with the Palo Alto Medical Foundation, a large multi-specialty physician group in California. The researchers develop and evaluate an online decision-support tool designed to help older adults choose a drug insurance plan. The tool uses a proprietary scoring technology from a third-party provider and assigns an expert score to each plan, which is a combination of an estimated total cost of the plan and the plan's "star rating." The study finds that the subjects who had access to the tool were more likely to select the plans suggested by it, with a more pronounced result for the group that had access to the expert scores in addition to the list of plans ordered by those scores. These results could help older adults make better drug insurance choices and improve the efficiency of the health sector. At the same time, the study shows that an algorithmic score can act as a coordinating device, similar to the findings of Paravisini and Schoar (2013). Although

other types of AI experiments could have found this coordinating role of a score, only a natural AI experiment provides an ultimate test that is directly relevant for the healthcare system.

All these examples illustrate why natural AI experiments can be the most fitting design choice. The use of AI technologies is complex and deeply embedded in organizational structures—culture, employee dynamics, specific business goals—something that controlled settings might struggle to replicate. AI usage in organizational settings can also reveal unanticipated behavioral responses and heterogeneity that are unlikely to emerge in simulated environments. Finally, the ethical and compliance considerations that affect AI deployment in organizations are context-dependent and are difficult to replicate in controlled settings, which would limit external validity.

3.5 Edge Cases

We chose the examples in the previous sections to highlight the clear-cut cases in each category. In this section, by contrast, we turn to the cases that may present a challenge for classification to stress-test the logic behind our taxonomy. We start with Dell'Acqua (2022) who studies how the quality of an algorithm affects workers' reliance on it and the resulting quality of decisions. The subjects in the study are freelance recruiters hired on an online platform to screen stylized job applications. The subjects' task is to choose whether to invite an applicant for an interview by trying to guess the applicant's math ability based on other characteristics, such as education and employment. The subjects are split into four treatment groups, depending on an algorithm they have access to. In the control group, subjects do not have access to an algorithm, while in the remaining three groups, subjects can either have an almost perfect algorithm, a "good" algorithm with 85% accuracy, or a "bad" algorithm with 75% accuracy. The algorithms are based on a related study (Cowgill et al., 2020) in which software engineers predict a person's math ability based on their education and employment characteristics. The study finds that subjects who have access to higher-quality AI are less accurate and spend less effort than subjects who have access to lower-quality AI, which suggests that a higher-quality AI may not always be beneficial for human decision-making.

This study is a field experiment, involving subjects in their natural roles, i.e., recruiters. It might appear, therefore, that this experiment belongs to a natural AI category. Our taxonomy, however, places it in the stylized AI category. The key consideration here is the implementation of AI. The algorithms used in the study are developed specifically for the purpose of a (related) study and are not intended to be used in actual hiring decisions.

Hence, the study does not fall into quasi-natural or natural AI categories, according to our definition. The study does implement an algorithm, hence, it is not a conceptual AI experiment either.

For another example, we consider Cox et al. (2016) who collaborate with a large U.S. hospital to study the uptake of a clinical decision-support system among resident physicians and fourth-year medical students. The decision-support system makes hospital discharge recommendations based on a probit model estimated on the data from the hospital's electronic medical records. In the experiment, subjects view patient charts from the database used to develop the decision-support system. They make discharge decisions based on either typical information or that information supplemented by a recommendation from the decision-support system. In the group that has access to the system, the default decision is either generated by the system or based on the current practices. The results show that the decision-support system is more effective if the default is generated by the system and subjects who override it have to explain their choice than if the information provided by the system is simply available.

It might appear that this study belongs to a conceptual AI category because subjects' choices are not consequential for actual patients. Subjects' choices only affect their own payoffs: they receive a monetary bonus for correct discharge decisions. Our taxonomy, however, places it in the quasi-natural AI category. The key consideration here, again, is the implementation of the algorithm. The study actually implements an algorithm—a decision-support system—hence, it is not a conceptual AI experiment according to our definition. The algorithm is developed to be ultimately used in an actual hospital and not for the purpose of conducting a study, hence, it is not a stylized AI experiment. Since researchers conduct a controlled experiment in artificial conditions, albeit with actual doctors, this is not a natural AI experiment. The researchers, in fact, justify their choice of a quasi-natural AI design over a natural one: "it is a practical and ethical requirement before application of the system on patient wards in hospitals" Cox et al. (2016)[P. 2].

Finally, we consider the case of Bai et al. (2022) who collaborate with a large Chinese warehouse operator owned by Alibaba, the largest retailer in China, to study the effects of human versus computer task assignment on workers' fairness perceptions and productivity. In the experiment, warehouse workers receive lists of items they need to pick from the warehouse (pick lists) either from a human supervisor or from a computer terminal. A list of items for each worker to pick is simply selected at random from a pool of available pick lists, previously generated by a logistical algorithm, regardless of whether a human supervisor or a computer distributes the list. Although the underlying rule for generating a pick list is

the same, the workers who receive pick lists from a computer perceive their assigned tasks as fairer than do workers who receive pick lists from a human supervisor. This effect results from a concern that a human supervisor might have been biased towards or against some workers. Greater fairness perception of the computer assignment translates into productivity gains of about 18%.

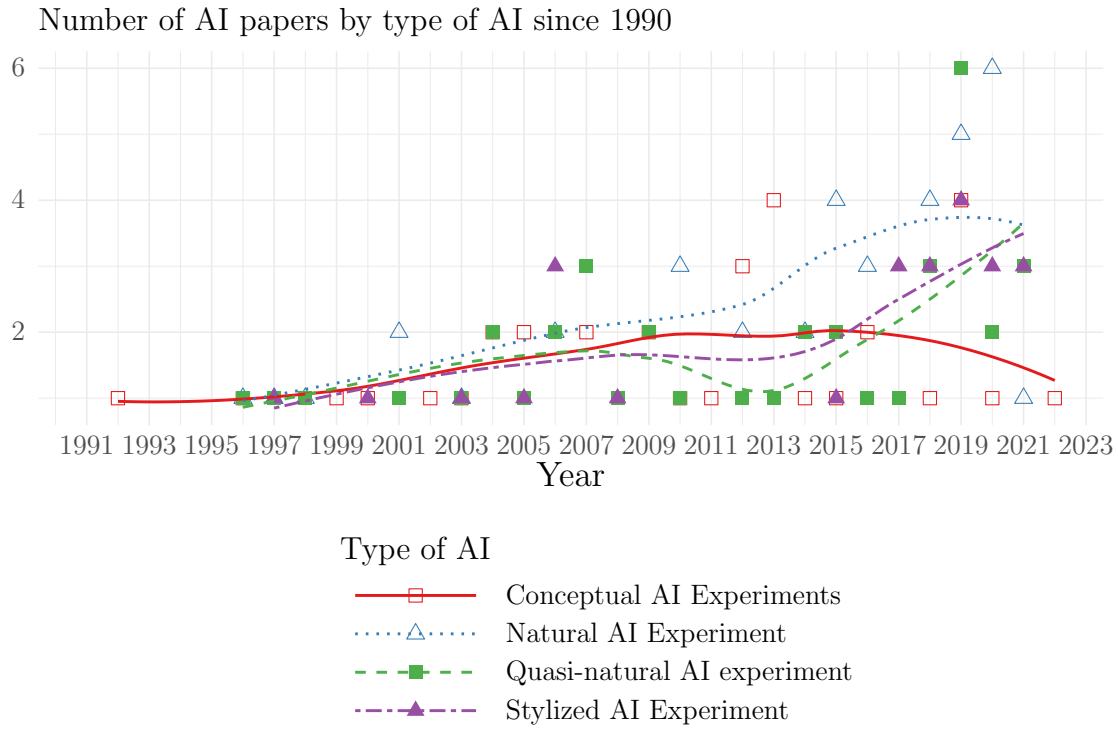
This study presents a challenge for classification because, while it is a field experiment, the algorithm for generating pick lists is a simple random assignment, there is no sophistication one would expect to see in a natural AI experiment. We think, however, that this study does belong to a natural AI category. The key consideration here is the purpose of designing an algorithm, even a simple one like random assignment: whether it is designed merely for the study or whether it will actually be implemented in a warehouse. The study is silent about that, however, given the high-stakes environment of a natural field experiment, it seems plausible that the algorithm is tested for the purpose of being deployed, perhaps with some modifications, in an actual work setting.

We conclude by noticing that our classification relies on how AI is implemented in a study, rather than on how a study is implemented, e.g., whether it is a field experiment or not. It follows that the same type of study, in principle, can be implemented using different AI experiments. Take, for example, correspondence/audit studies (Verhaeghe, 2022) in which researchers typically send fictional CVs of potential job applicants to recruiters. Researchers can implement a correspondence study as a conceptual AI experiment, e.g., if they vary whether an applicant has experience with generating prompts for LLMs. Alternatively, they can implement it as a natural AI experiment, if they vary whether they send job applications to recruiters who are known to use algorithms for screening candidates or not.

4 Discussion

We have presented the four types of AI experiments in a neutral manner, balancing their strengths with their weaknesses. In practice, however, these types of AI experiments vary in prevalence, as illustrated in Figure 3. All four types experienced increased usage since the 1990s, however, after the early 2010s, their trajectories began to diverge. The popularity of conceptual AI experiments peaked around mid-2010 and has subsequently declined. Stylized and quasi-natural AI experiments initially decreased in popularity during the mid-2010s but have shown a resurgence in recent years. Natural AI experiments have enjoyed a steady growth throughout the observation period.

Figure 3: Popularity of AI experiments by type


Number of AI papers by type of AI since 1990
6
4
2
1991 1993 1995 1997 1999 2001 2003 2005 2007 2009 2011 2013 2015 2017 2019 2021 2023
Year
Type of AI
Conceptual AI Experiments
Natural AI Experiment
Quasi-natural AI experiment
Stylized AI Experiment

Note: The figure shows the total number of experimental AI papers published in a given year since 1990, clustered by type. The line shows a loess fit. The list of papers is based on Chugunova and Sele (2022) and March (2021).

The strengths and weaknesses of the four types of AI experiments that we identify offer insights into possible reasons behind these trends. Conceptual AI experiments, while valuable for rapid generation of new insights, likely saw a decline as researchers sought to validate initial findings with experiments where AI is actually implemented. The relative ease of implementation, combined with their ability to model dynamic interactions beyond what simple vignettes allow, has likely contributed to the renewed interest in stylized and quasi-natural AI experiments. The increasing accessibility and capabilities of LLMs have likely further fueled interest in quasi-natural AI experiments (Charness, Jabarian, and List, 2025). A consistent growth in natural AI experiments likely reflects the recent wave of AI adoption and automation within organizations (Agrawal, Gans, and Goldfarb, 2022).

Based on these trends, we anticipate continued growth in AI experiments that implement AI in some form. The proliferation of LLMs and the ability to fine-tune them will likely drive further adoption of quasi-natural AI experiments, enabling researchers to investigate interactions with sophisticated conversational agents (Leib et al., 2023). Ongoing adoption of AI

tools and automation into workplaces will likely stimulate the use of natural AI experiments, e.g., to study the productivity effects of new technologies (Dell'Acqua et al., 2023; Brynjolfsson, Li, and Raymond, 2025). Stylized AI experiments will likely still find place in future research agendas. These experiments enable researchers to explore underlying behavioral mechanisms by controlling how an algorithm is constructed and presented, which distinguishes them from the largely "black box" nature of LLMs. Moreover, while LLMs, at least currently, are best suited for research questions involving textual data and conversational interactions, stylized AI experiments can address a broader range of questions. It is conceivable, however, that the heightened interest in LLMs will inspire new research questions that can be answered with this tool (Charness, Jabarian, and List, 2025; Korinek, 2023). Finally, conceptual AI experiments may assume a complementary role, supplementing the findings from the other types of AI experiments.

5 Conclusion

In this paper, we propose a taxonomy of AI experiments. Our taxonomy features four types of AI experiments: conceptual AI experiments, stylized AI experiments, quasi-natural AI experiments, and natural AI experiments. At the core of our taxonomy is the sophistication of AI used. To evaluate the sophistication, we propose a simple and robust proxy test of whether AI is developed exclusively for a research study. We provide a guide on the advantages, disadvantages, and best use cases for each type, illustrated via various examples.

Although we designed our taxonomy to be easy-to-use and robust, we recognize that there might be experiments that are difficult to classify using our procedure. We illustrate a few of such edge cases to stress-test our classification. New types of AI experiments may also emerge in the future that would warrant an update of the taxonomy. We believe, however, that our taxonomy will prove to be a useful tool for organizing the existing literature and will help researchers design new experiments.

References

Agrawal, A., J. Gans, and A. Goldfarb. 2022. "ChatGPT and how AI disrupts industries." *Harvard Business Review* 12:1–6.

Alekseev, A. 2025. "The Economics of Babysitting a Robot." Working paper, University of Regensburg.

Anelli, M., I. Colantone, and P. Stanig. 2019. "We Were The Robots: Automation and Voting Behavior in Western Europe." CReAM Discussion Paper Series No. 17/19, London.

Angerer, M., T. Neugebauer, and J. Shachat. 2023. "Arbitrage bots in experimental asset markets." *Journal of Economic Behavior & Organization* 206:262–278.

Awad, E., S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.F. Bonnefon, and I. Rahwan. 2018. "The Moral Machine experiment." *Nature* 563:59–64.

Azevedo, E.M., A. Deng, J.L. Montiel Olea, J. Rao, and E.G. Weyl. 2020. "A/B Testing with Fat Tails." *Journal of Political Economy* 128:4614–000.

Bai, B., H. Dai, D.J. Zhang, F. Zhang, and H. Hu. 2022. "The Impacts of Algorithmic Work Assignment on Fairness Perceptions and Productivity: Evidence from Field Experiments." *Manufacturing & Service Operations Management* 24:3060–3078.

Bao, T., E. Nekrasova, T. Neugebauer, and Y.E. Riyanto. 2022. *Chapter 23: Algorithmic trading in experimental markets with human traders: A literature survey*, Cheltenham, UK: Edward Elgar Publishing. pp. 302–322.

Brynjolfsson, E., D. Li, and L. Raymond. 2025. "Generative AI at Work." *The Quarterly Journal of Economics* 140:889—942.

Bundorf, M.K., M. Polyakova, and M. Tai-Seale. 2019. "How do Humans Interact with Algorithms? Experimental Evidence from Health Insurance." Working Paper No. 25976, National Bureau of Economic Research.

Burton, J.W., M.K. Stein, and T.B. Jensen. 2020. "A systematic review of algorithm aversion in augmented decision making." *Journal of Behavioral Decision Making* 33:220–239.

Cappelen, A.W., C. Cappelen, and B. Tungodden. 2023. "Second-Best Fairness: The Trade-off between False Positives and False Negatives." *American Economic Review* 113:2458–85.

Carvajal, D., C. Franco, and S. Isaksson. 2024. "Will Artificial Intelligence Get in the Way of Achieving Gender Equality?" *SSRN Electronic Journal*, pp. .

Castelo, N., M.W. Bos, and D.R. Lehmann. 2019. "Task-Dependent Algorithm Aversion." *Journal of Marketing Research* 56:809–825.

Charness, G., B. Jabarian, and J.A. List. 2025. "The next generation of experimental research with LLMs." *Nature Human Behaviour*, pp. .

Chugunova, M., D. Harhoff, K. Hölzle, V. Kaschub, S. Malagimani, U. Morgalla, and R. Rose. 2025. "Who Uses AI in Research, and for What? Large-scale Survey Evidence from Germany." Research Paper No. 25-11, Max Planck Institute for Innovation & Competition.

Chugunova, M., and D. Sele. 2022. "We and It: An interdisciplinary review of the experimental evidence on how humans interact with machines." *Journal of Behavioral and Experimental Economics* 99:101897.

Cominelli, L., F. Feri, R. Garofalo, C. Giannetti, M.A. Meléndez-Jiménez, A. Greco, M. Nardelli, E.P. Scilingo, and O. Kirchkamp. 2021. "Promises and trust in human–robot interaction." *Scientific Reports* 11:9687.

Corgnet, B., R. Hernán-González, and R. Mateo. 2023. "Peer effects in an automated world." *Labour Economics* 85:102455.

Cowgill, B., F. Dell'Acqua, S. Deng, D. Hsu, N. Verma, and A. Chaintreau. 2020. "Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics." In *Proceedings of the 21st ACM Conference on Economics and Computation*. New York, NY, USA: Association for Computing Machinery, EC '20, pp. 679–681.

Cox, J.C., V. Sadiraj, K.E. Schnier, and J.F. Sweeney. 2016. "Higher quality and lower cost from improving hospital discharge decision making." *Journal of Economic Behavior & Organization* 131:1–16.

Dargnies, M.P., R. Hakimov, and D. Kübler. 2024. "Aversion to Hiring Algorithms: Transparency, Gender Profiling, and Self-Confidence." *Management Science*, pp. .

Dell'Acqua, F. 2022. "Falling asleep at the wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters." Working paper, Harvard Business School.

Dell'Acqua, F., E. McFowland III, E.R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K.R. Lakhani. 2023. "Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality." Working Paper No. 24-013, Harvard Business School.

Dietvorst, B.J., J.P. Simmons, and C. Massey. 2015. "Algorithm aversion: People erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology: General* 144:114–126.

Falk, A., T. Neuber, and N. Szech. 2020. "Diffusion of Being Pivotal and Immoral Outcomes." *The Review of Economic Studies* 87:2205–2229.

Farjam, M., and O. Kirchkamp. 2018. "Bubbles in hybrid markets: How expectations about algorithmic trading affect human trading." *Journal of Economic Behavior & Organization* 146:248–269.

Gallego, A., A. Kuo, D. Manzano, and J. Fernández-Albertos. 2022. "Technological Risk and Policy Preferences." *Comparative Political Studies* 55:60–92.

Gogoll, J., and M. Uhl. 2018. "Rage against the machine: Automation in the moral domain." *Journal of Behavioral and Experimental Economics* 74:97–103.

Gorny, P.M., B. Renner, and L. Schäfer. 2023. "Prosocial behavior among human workers in robot-augmented production teams—A field-in-the-lab experiment." *Frontiers in Behavioral Economics* 2.

Granulo, A., C. Fuchs, and S. Puntoni. 2021. "Preference for human (vs. robotic) labor is stronger in symbolic consumption contexts." *Journal of Consumer Psychology* 31:72–80.

Harrison, G.W., and J.A. List. 2004. "Field experiments." *Journal of Economic Literature* 42:1009–1055.

Hergueux, J., and N. Jacquemet. 2015. "Social preferences in the online laboratory: a randomized experiment." *Experimental Economics* 18:251–283.

Jacob Leal, S., and N. Hanaki. 2023. "Algorithmic trading, what if it is just an illusion? Evidence from experimental asset markets." Working paper.

Jeffrey, K. 2021. "Automation and the future of work: How rhetoric shapes the response in policy preferences." *Journal of Economic Behavior & Organization* 192:417–433.

Jussupow, E., I. Benbasat, and A. Heinzl. 2020. "Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion." Research Paper No. 168, AIS Electronic Library.

Kirchkamp, O., and C. Strobel. 2019. "Sharing responsibility with a machine." *Journal of Behavioral and Experimental Economics* 80:25–33.

Klockmann, V., A. von Schenk, and M.C. Villeval. 2022. "Artificial intelligence, ethics, and intergenerational responsibility." *Journal of Economic Behavior & Organization* 203:284–317.

Korinek, A. 2023. "Generative AI for Economic Research: Use Cases and Implications for Economists." *Journal of Economic Literature* 61:1281–1317.

Langer, M., C.J. König, and V. Hemsing. 2020. "Is anybody listening? The impact of automatically evaluated job interviews on impression management and applicant reactions." *Journal of Managerial Psychology* 35:271–284.

Langer, M., and R.N. Landers. 2021. "The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers." *Computers in Human Behavior* 123:106878.

Lee, M.K. 2018. "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management." *Big Data & Society* 5:2053951718756684.

Leib, M., N. Köbis, R.M. Rilke, M. Hagens, and B. Irlenbusch. 2023. "Corrupted by Algorithms? How AI-generated and Human-written Advice Shape (Dis)honesty." *The Economic Journal* 134:766–784.

Luo, X., S. Tong, Z. Fang, and Z. Qu. 2019. "Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases." *Marketing Science* 38:937–947.

March, C. 2021. "Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players." *Journal of Economic Psychology* 87:102426.

Meehl, P.E. 1954. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press.

Palan, S., and C. Schitter. 2018. "Prolific.ac—A subject pool for online experiments." *Journal of Behavioral and Experimental Finance* 17:22–27.

Paravisini, D., and A. Schoar. 2013. "The Incentive Effect of Scores: Randomized Evidence from Credit Committees." Working Paper No. 19303, National Bureau of Economic Research, August.

Pioggia, G., A. Ahluwalia, F. Carpi, A. Marchetti, M. Ferro, W. Rocchia, and D.D. Rossi. 2004. "FACE: facial automaton for conveying emotions." *Applied Bionics and Biomechanics* 1:91–100.

Russell, S.J., and P. Norvig. 2021. *Artificial intelligence: A modern approach*, 4th ed. Pearson.

Schauer, A., and D. Schnurr. 2023. "Competition and Collaboration between Human and Artificial Intelligence in Digital Markets." Working paper.

Strobel, C. 2025. "The impact of process automation on performance." *Journal of Behavioral and Experimental Economics*, pp. 102377.

Verhaeghe, P.P. 2022. *Correspondence Studies*, Cham: Springer. pp. 1–19.

Wang, B., and A. Komatsuzaki. 2021. "GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model."

Webb, M. 2019. "The Impact of Artificial Intelligence on the Labor Market." Working paper.

Werner, T. 2021. "Algorithmic and Human Collusion." Working paper.

Wu, N. 2022. "Misattributed blame? Attitudes toward globalization in the age of automation." *Political Science Research and Methods* 10:470–487.

Yeomans, M., A. Shah, S. Mullainathan, and J. Kleinberg. 2019. "Making sense of recommendations." *Journal of Behavioral Decision Making* 32:403–414.

Zhang, B. 2022. "No Rage Against the Machines: Threat of Automation Does Not Change Policy Preferences." In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery, AIES '22, pp. 856–866.